A 90 nm CMOS, 6 µW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection

Komail Badami, Steven Lauwereins, Wannes Meert, Marian Verhelst, KU Leuven, Belgium

Abstract – This work presents a sub-6 µW acoustic front-end for speech/non-speech classification in a voice activity detection (VAD) in 90 nm CMOS. Power consumption of the VAD system is minimized by architectural design around a new Power-Proportional sensing paradigm and the use of machine-learning assisted moderate-precision analog analytics for classification. Power-Proportional sensing allows for hierarchical and context-aware scaling of the frontend's power consumption depending on the complexity of the ongoing information extraction, while the use of analog analytics brings increased power efficiency through switching on/off the computation of individual features depending on the features' usefulness in a particular context. The proposed VAD system reduces the power consumption by 10X as compared to state-of-the-art systems and yet achieves an 89% average hit rate for a 12 dB signal to acoustic noise ratio in babble context, which is at par with software based VAD systems.

I. INTRODUCTION

Technological innovations are changing the way we interact with electronic devices. Interactions like voice control and gesture recognition are rapidly gaining popularity. Such natural interactive systems do not only need many integrated sensors, but also always-awake, reactive sensor frontends. These frontends generate large amounts of raw signals that state-of-the art (SotA) frontends immediately digitize for processing on a DSP. This very robust approach is

not power efficient, as not all raw sensor signals are equally relevant. The net information content of a sensed signal is quite often significantly smaller than the Nyquist rate [1-7]. Existing works such as Information-Rate processing [1,2], Analog to Information conversion [3-5] and Compressed Sensing [6,7] show power savings by extracting or compressing the information from signals before digitizing the data. However, as these schemes operate in a static way, the compression or extraction parameters are set beforehand. Yet, the information content in raw signals and its application relevance dynamically varies depending on the operating context. Operating such systems efficiently hence requires a dynamic system adaptation depending on the context or signal information content. Existing systems do not perform such fine grain adaptive behavior, which severely limits their power savings as shown by solid line in Fig. 1.

We propose a self-scalable, Power-Proportional sensing paradigm which gracefully scales the system's power consumption with the amount and complexity of extracted information, i.e. the power consumption for such a system increases only as the task of information extraction gets more complex. To this end, in this paper we propose key enablers for Power-Proportionality and apply them to a proof of concept acoustic frontend for voice activity detection (VAD).

VAD systems distinguish speech from non-speech in different background noise contexts for varying signal to acoustic noise ratios (SANR). SotA VAD systems [8-10] extract complex features like Mel-Frequency Cepstral Coefficients, DCT etc. to differentiate speech from non-speech. The high computational complexity of such features results in large power consumption, typically about 50 - 100 µW [8-11] in addition to the power consumption of the required active microphone. Such a continuous large power consumption is unacceptable for battery powered always-on sensor frontends. This work exploits our new Power-Proportional sensing paradigm along with moderate-precision, computationally-inexpensive, analog feature-extraction, coupled

with an embedded mixed-signal classifier to save more than 10X power consumption over SotA without compromising on the classification accuracy.

The outline for this paper is as follows. Section II discusses insights into the design principles for Power-Proportional sensing and explains the rationale behind the analog feature-extraction instead of the commonly used digital scheme. Section III describes the architecture and specification set for VAD while the detailed implementation is discussed in Section IV. Measurement results for the chip and for the full VAD system are discussed in Section V.

II. KEY PRINCIPLES FOR POWER EFFICIENT SENSING

This section details the two key principles that allow our always-on sensing system to scale its power consumption with the information extracted saving 10X power over SotA VAD systems.

A. Power-Proportional Sensing

The core premise for Power-Proportional sensing is that power consumption of the sensing system scales proportionally with the complexity of the sensing task. The sensing process with the target of information extraction can increase in complexity along two dimensions:

First, the amount of information extracted from the incoming signal can scale in complexity. Consider for example, the task of speaker identification v/s speech detection. The former task entails the later as a prerequisite first step, hence justifying the increase in power consumption. Enabling hierarchical operation for tasks of increasing complexity allows scaling of power consumption with complexity of information extraction. In such an architecture each processing stage extracts more complex information than the previous stage while consuming more power. This enables information extraction by necessity, as is shown on the horizontal-axis in Fig. 1.

Secondly, even if the amount of extracted information remains the same, distinguishing the useful information from the background noise (the context) is subject to varying levels of

difficulty. For this case consider the complexity of speech detection in a quiet office, in contrast to a noisy street environment. The amount of information needed is same in both cases, but in the latter case as the background noise maps directly onto the information spectrum, it creates in-band interference on the desired signal. As such, distinguishing speech from non-speech becomes more complex, hence justifying the increase in power consumption. Context-awareness enables Power-Proportional sensing to scale power as the background noise context scales the complexity of information extraction, as shown in bold in Fig. 1. For the example above, context-awareness allows to use a much smaller discriminating feature subset in a low noise environment and a relatively larger subset for noisy background contexts, hence scaling power.

SotA sensing systems do not exploit the power scaling opportunity offered by the above scenarios, and typically operate constantly in full processing mode. This plateaus the on-state power consumption for SotA sensing systems independent of system utility as shown in Fig. 1.

B. Power Efficiency through Analog Analytics

The Power-Proportional sensing paradigm as highlighted in previous paragraph needs complexity and precision dependent power scalable hardware blocks. Such power scaling with precision is very different for analog and digital implementations. Analog power consumption scales gradually for thermal noise limited system with low-to-medium precision, while digital has a logarithmic power v/s precision profile. As it has been shown in [12] and in Fig. 2, for a 0.25 µm CMOS technology, analog computation is not only more power-efficient than digital for low-to-medium resolution processing, but also exhibits better scalability.

Reduction in supply voltage due to technology scaling allows more power efficient digital circuits and questions the beneficial analog behavior in advanced technologies. This is because with scaling, the cost of maintaining the same precision in analog increases as a larger bias

current is needed to reduce the noise-floor compensating for reduction in signal swing. Assuming that the supply voltage has scaled from 2.5 V for 0.25 µm to 0.9 V for a 40 nm technology, the active digital power has scaled down by $10\log(2.5^2/0.9^2)$ ~ 9 dB while analog power consumption goes up by 4.5 dB [12] for subthreshold design. Contrasting effects of reduction in average capacitance per node and increase in subthreshold-leakage on digital power consumption are not considered here. The above discussion implies that while analog keeps its favorable scalability, the analog-digital efficiency crossover point moves towards the left by 2 bits. This renders analog computation cheaper than digital for up to 7 bits of precision as shown in Fig. 2.

Digital enhancements, such as machine learning and calibration, can restore some of the lost benefit of analog over digital computation for always-on sensing or classification tasks because these often do not need perfect signal reconstruction, but only need error resilient processing such as detection or classification. Specifically such tasks do not require accurate absolute computations, but only relative comparisons of the computed feature values to on-chip trained thresholds, as we will show in the design presented in this paper. Hence, absolute precision requirements for such systems are rather modest, and mismatches and offset impairments are automatically taken care of by the embedded trained classifier in the loop. As demonstrated by this work, as well as some existing works, machine learning assisted [13, 14] and/or digital calibration [15] can improve SNR by 6 – 10 dB for comparable power which pushes the efficiency crossover point in the rightward direction as shown in Fig. 2. These estimations support the use of analog computation for systems requiring scalability up to 8 bits of precision.

III. SYSTEM ARCHITECTURE AND SPECIFICATIONS

This section highlights the use of the aforementioned key principles in the developed VAD architecture [16] and derives the specifications for the analog/mixed-signal building blocks.

A. VAD System Architecture

The top-level block diagram of the proposed Power-Proportional VAD system is shown in Fig. 3. The main sub-blocks of the system are the threshold based wakeup block, the analog feature-extractor, the mixed-signal classifier and the microcontroller, which operate in the described Power-Proportional sensing fashion as follows:

An always-awake threshold-based wakeup block keeps checking the passive microphone for sound activity. When any signal – not necessarily useful – is detected, it wakes up the analog feature-extractor that translates the input signal into a set of features. The on-chip classifier uses these computed features to classify the incoming signal as speech/non-speech. If the signal is speech, the classifier wakes up the microcontroller for more advanced processing.

Such hierarchical activation of information extraction hardware allows the VAD system to be in the lowest power-mode possible, while still able to compute the necessary information. This allows scaling the power with necessary information as outlined in Section II.A.1. Also, as not all computed analog features carry information under all background noise contexts, machine learning based context-awareness allows dynamically disabling the computation of features that do not assist in classification. Such context-aware computing allows further power scaling depending on the number of useful features necessary as explained in Section II.A.2. The control of feature activation and classifier configuration is done by the embedded micro-controller. This microcontroller periodically wakes up to check for background noise context-changes and upon detecting a change, retrains the classifier and activates the required features for the new context. As further modelled in subsection B, considering that the analog feature-extraction blocks are in the loop during this training operation, all static analog impairments such as mismatch, gain errors, or offsets are absorbed in the trained feature thresholds and do not affect the classification

accuracy. This justifies the usage of low-precision analog analytics for feature computation, as discussed in Section II.B. Before detailing the design of individual sub-blocks in Section IV, subsection III.B derives specifications for the targeted VAD system.

B. Specifications for VAD system

This section first derives the system level specifications and then the specifications for individual analog blocks. The system computes an analog feature-set for the acoustic signal by decomposing the signal into different frequency bands and then extracting the average value of the rectified signal in each frequency band. Mathematically, each analog feature af_i is defined as

$$af_i = \overline{abs[Ax(t) * h_i^{BPF}]} \tag{1}$$

where $Ax(t)$ is the amplified acoustic signal, h_i^{BPF} is the impulse response of band pass filter used to decompose the input signal into a smaller frequency band, abs,$*$ and $\overline{x}$ represent the absolute value, convolution, and averaging respectively. The features hence represent the average power present in every frequency band. It is therefore important to determine the required frequency range, number of observed frequency bands, and the necessary precision, as these parameters will strongly influence the classification accuracy as well as the system's power consumption. Such system-specifications are evaluated based on a MATLAB model of the analog feature-extractor of VAD system based on equation (1).

Along the frequency axis, the bulk of energy for speech and acoustic noise is concentrated in the frequency range 100 Hz – 4 kHz [17]. The MATLAB model varies the number of computed features in the above frequency range by scaling the Q factor of the band pass filters. This ensures that the entire frequency range is always populated with filters, with an increasing frequency resolution as the number of computed features increases. The results of the above simulation are shown in Fig. 4(a). It can be seen that more features improve classification

accuracy, yet accuracy gains diminish beyond 16 features allowing us to limit our design to a maximum of 16 (individually (dis)activated) features. Further, the model also evaluates the impact of static analog impairments, for example by degrading the gain in the signal path, as seen in Fig. 4(b), as long as these occur within the training loop, they are absorbed in the thresholds learnt for classification and hence have no impact on classification accuracy.

Fig. 5 histogram shows the relative relevance of each of the 16 analog features in the speech v/s non-speech classification for exhibition noise context with 0 dB SANR. It is clear that the middle-frequency features af_5 to af_{12} are more commonly used. Hence we only pass these features to an on-chip classifier, while the full feature-set is passed on to a microcontroller only when needed for more complex tasks, such as context-change detection.

Another important group of parameters are the maximum input-referred noise and the necessary gain for the system. The specifications for input-referred noise and gain strongly depend on the input signal level, which depend on the type and make of the microphones used in the system. The active microphones used by SotA VADs consume 20 - 50 µW [18, 19] in addition to the power consumption of the VAD circuitry itself. This is unacceptably high for always-on sensing acoustic systems. Such systems hence necessitate the use of passive microphones in low power budget applications. Such passive microphones typically have a sensitivity down to – 60 dBV. This translates to an rms signal level of 30 µV @ 65 dB sound pressure level (SPL) for a nominal conversation at 1 m distance [20]. This limits the maximum allowable noise-floor to less than 30 µVrms and also decides the minimum gain necessary in the amplifier depending on the LSB size, being 45 dB to achieve 8 bit precision over 1V. This design has a gain-range from 20 to 80 dB in 20 dB steps to cover a wide range of input signals although we anticipate that only up to 60 dB would be necessary. Also the averaging time

depends on the frequency of classification which in a typical VAD system is every 10-16ms [8-10]. This averaging is implemented as LPF with a f_{-3dB} of 16 Hz. A summary of the VAD system-specifications is highlighted in Table 1.

IV. SYSTEM IMPLEMENTATION

This Section details the implementation nuances of the individual system blocks discussed in the previous section: namely the wakeup detector, the analog feature-extractor and the embedded mixed-signal classifier. A further subsection discusses system training for the complete VAD system before discussing one-time calibration and measurement results in Section V.

A. Wakeup detector

The always-awake threshold-based wakeup detector acts as the system's watch-dog that wakes up the analog feature-extractor only when a signal of sufficient strength is detected. A single bit of information indicating presence or absence of acoustic signal is needed. The wakeup detector is a low power 3-phase comparator and its schematic is shown in Fig. 6. As the input signal level for this comparator can be as low as 30 µV and the comparator reference $Vref_{comp}$, is generated using 1.2 V, 8-bit DAC, at least 45 dB gain is necessary in the pre-amplifier to keep the signal swing greater than 1 LSB ~ 4.5 mV.

The preamplifier is a cascade of four single stage amplifiers. Each amplifier is a PMOS input source-coupled single-ended differential amplifier and can be turned on/off individually to save power depending on the microphone's signal-level and is designed to provide a mid-band gain of 20 dB. The f_{-3dB} of the amplifier is limited to 2 kHz as only the speech envelope needs to be detected. The comparator $Vref_{comp}$ can potentially vary as per the ambient noise-level, but this is beyond the scope of this work. Measured power consumption of this block is 700 nW when all four amplifier stages are turned on, and excluding the external bias.

B. Analog feature-extractor

On receiving the wakeup signal from the threshold based wakeup detector, the analog feature-extractor decomposes the input signal into the set of 16 features. The on-chip classifier evaluates whether the signal is potentially speech or background noise by comparing a feature subset to trained thresholds in a Decision Tree (DT) topology (see subsection C). This subsection first describes the flow of the acoustic signal through the analog feature-extractor, followed by the implementation details of the individual blocks that participate in feature-extraction.

Fig. 7 shows the detailed architecture for the analog feature-extractor. The signal from the passive microphone after low noise amplification is fed to 16 bands. Each band allows further amplification and does a BPF operation with exponentially spaced f_c to mimic human hearing [21]. The output of each BPF filter is averaged by a rectification and LPF operation which results in 16 analog features af_1 - af_{16}, from which the subset $af_5 – af_{12}$ is used by the on-chip classifier.

The partitioning of the amplification between the shared LNA and the individual frequency bands allows a finer control over necessary amplification in each band. This contributes to Power-Proportional information extraction, as it allows turning off amplifier stages of unused features along with all other circuitry involved in individual feature computation. This enables context-aware power savings, as discussed in Section II.A.2. The sub-blocks of the analog feature-extractor are now explained in more detail.

1) LNA & Amplifiers

The LNA is interfaced with a passive microphone and is designed to provide a mid-band gain of 20 dB up to a frequency range of 5 kHz while keeping the rms integrated input-referred noise smaller than 30 µV. The LNA is shared across all 16 bands as can be seen from Fig 7. Further amplification in each band is done through a cascade of four individually controllable single

stage amplifiers with each stage designed to provide 20 dB gain as in Fig. 7. A single stage amplifier topology was chosen for both LNA and in-band amplifiers for efficiency reasons, to avoid the power overhead of pushing non-dominant pole(s) beyond the unity gain bandwidth. The closed loop gain error introduced due to insufficient open loop gain is a static error and is, as discussed, absorbed in the training phase.

The pseudo resistive feedback fixes the output bias point of the amplifier as shown in Fig 8. As the area for the input transistors is large (80 μm X 10 μm) to reduce the flicker noise, gate leakage current up to 20 pA can shift the output bias point by as much as 50 mV due to voltage drop across the pseudo resistor. The inter-stage capacitive coupling however ensures the bias point shift is not cascaded to next stage.

As discussed next, the band pass filters across the bands have increasing center frequencies. To cover for this, the f_{-3dB} of the amplifiers in each band also increases progressively from band 1 to band 16. This is illustrated by the simulated magnitude response of the amplifiers in Fig. 9.

2) Band Pass Filters

The amplifier output in each of the 16 bands is passed through a band pass filter (BPF) whose center frequency (f_c) increases exponentially from 75 Hz in band 1 to 5 kHz in band 16. The f_c for a second order gm–C filter (see Fig. 10) is scaled by varying the bias current across the bands. From the BPF frequency response in Fig. 11 it can be seen that stop-band attenuation for individual filters is better than –40 dB but the adjacent band rejection is only –1.5 dB. This adds redundancy in the extracted features, leading to a high correlation between features of adjacent channels. This makes the system tolerant to shifts in the center frequency of BPFs.

3) Averaging circuit

The output of each BPF is averaged individually by first rectifying and then low-pass filtering

with a f_{-3dB} of 16 Hz to result in 16 analog features ($af_1 - af_{16}$). The architecture of the current-mode averaging is shown in Fig. 12. Normally-off transistors used for rectification (in dotted box) turn on based on the direction of current from the BPF. The current steering network makes the current direction unipolar and is read across the gm-based resistors. A first order gm - C LPF extracts the average value of this unipolar signal. Such normally-off transistors result in asymmetric rectification (dashed line) as in Fig. 13. This adds a dc-offset to the computed feature level shown by the averaged line (dot - dashed) in Fig. 13. Such offsets can be learnt during the training phase and do not affect classification accuracy.

C. Decision tree based classifier

The extracted feature subset, af_5 - af_{12}, is passed on to the on-chip classifier (Fig. 5) while the complete feature-set af_1 - af_{16} can be passed on to an off-chip ADC for more complex information extraction, such as context-change detection and retraining the classifier as in [22]. In these cases, the Nyquist sampling rate for the features is only 16x2x16 = 512 Hz instead of 8 kHz for audio. The external ADC is *not* needed for embedded speech/non-speech classification.

The implementation of the on-chip 7-node 3-level mixed-signal Decision Tree classifier is shown in Fig. 14. Each node of the decision tree can be configured to select one feature out of af_5 - af_{12}. The selected feature (sf_i) is then compared with a reference voltage ($Vref_i$) determined by a modified C4.5 machine learning algorithm [22], generating the output decision b_i of each node:

$$b_i = xor[(sf_i > Vref_i), inv_i] \quad (2)$$

where inv_i bit sets the comparison direction. The decision fusion logic shown in Fig. 14 combines the outputs of all decision tree nodes.

D. VAD System training

The decision tree configuration and the individual feature activation is done using machine

learning which selects the most discriminative features between speech and the current background noise context. To this end, the on-chip decision tree classifier is trained with our modified C4.5 algorithm with 160 s of labeled data from the standardized NOIZEUS database [23]. The traditional C4.5 algorithm selects a feature-set to maximize the total *information-gain*. Our modification to C4.5 maximizes the *information-gain/watt* and therefore outputs a *resource-efficient* model that maximizes the information capture while minimizing the power [22]. This is enabled as each feature extracts information from a higher frequency band so that the power cost increases from af_1 to af_{16}. This maximization of *information-gain/watt* furthers Power-Proportionality by increasing power consumption only for more (complex) information. The training runs on the microcontroller to generate a discriminating feature subset and reference levels for the comparison in the DT. The training results of the past context are not stored but dynamically learnt, as context-change is detected [22].

V. MEASUREMENT SETUP AND RESULTS

The proposed system has been implemented on a 2 mm^2 chip in 90 nm CMOS as shown in Fig. 15. This section details the measurement results for the chip and for complete VAD system.

A. Chip performance results

This subsection first discusses the measurement results for the LNA and some individual blocks in the 16^{th} feature band in the chip followed by measurement results for complete bands.

The input-referred noise for the LNA is shown in Fig. 16. The noise has been measured at the LNA output over a frequency range of 10 Hz to 10 kHz. The rms input-referred integrated noise over the range of 75 Hz to 10 kHz is 32.5 µV. The total input-referred noise is expected to be 15% larger as this does not include the contributions from subsequent amplifier stages. For 3% and 5% THD, dynamic range is measured to be 40.2 dB and 45.4 dB respectively at 1 kHz.

Frequency responses of the individual blocks in the 16th feature band are shown in Fig. 17. Compared to simulation the mid-band gain of the LNA is reduced by 4 dB, which is estimated to be due to insufficient open loop gain. The large signal frequency response for the complete bands is shown in Fig. 18. As the f_c of the BPFs increase across bands, each band progressively processes higher frequency content to compute a feature, hence for a constant capacitive load the power consumption increases from band 1 to band 16 as it can be seen from Fig. 19. As already mentioned in section IV (D) this allows a power-aware learning to enable efficient classification. Finally, the measured rms noise at the output of each band is less than 2 mV. For an output signal range of 400 mV, this gives 7.5 bits of precision.

B. System measurement results

The chip is integrated with the microcontroller using external level-shifters and DACs, to form the complete VAD. Fig. 20 shows a one-time calibration to characterize for mismatch in the ADC and DAC paths. This subsection also displays the classification accuracy results for the complete VAD system and illustrates the achieved Power-Proportionality.

Receiver operating characteristic (ROC) curves characterize the classifier systems and depict hit-rates (HR) for the variables under observation [24]. Fig. 21 ROC curve shows that classification accuracy of our on-chip classifier is on-par with software based VAD systems of [8, 9, 25]. Further Fig. 22 validates the classification capacity over multiple contexts with different background noise conditions. Table 2 illustrates the Power-Proportional sensing in our VAD system by showing the gradual increase in system power consumption with the sensing task complexity. The power consumption for signal detection is measured to be below 1 µW, whereas power consumption for classification varies depending on complexity of the operating context and has an upper bound of 6 µW. The power consumption for background context-

change detection and relearning the DT is estimated to be 57 µW on a cortex M4 microcontroller. It is predicted that the VAD system will be 80% of the time in detection mode, 15% in classification mode and about 5% of time performing complex tasks, such as re-learning the context or decision tree training. The resulting duty-cycled power consumption is 3.8 µW for babble noise context. Further, the estimated power overhead for generating on-chip (currently off-chip) reference voltages for the comparators is leakage limited and is estimated to be less than 50 nW per reference value [26] as the reference voltage needs to drive only the gate nodes at near dc speed. Table 3 compares our work to SotA VADs [8-10, 25] and similar systems [27]. While maintaining the same classification accuracy as compared to software VADs, our system reduces the power consumption by 10X. Although hierarchical information extraction adds a maximum latency of 100 ms to the VAD decision task, this does not cause significant information loss as this latency is smaller than the average duration of a spoken vowel [28].

VI. CONCLUSIONS

This work demonstrates a power efficient acoustic sensing frontend for speech/non-speech classification in a voice activity detection system. The power efficiency is achieved by the use of machine learning assisted analog feature computation and by infusing the Power-Proportionality paradigm in various ways throughout the architecture. The use of analog features for information extraction allows individual turning on/off of features depending on the usefulness of a feature in a particular context while the Power-Proportionality concept controls the hierarchical activation of different sub-blocks depending on the complexity of the information extraction task. The idea of Power-Proportional sensing is demonstrated for an acoustic sensing system and can be extended to other systems such as motion and image sensing systems.

References

1. E-Hung Chen; Leven, William; Warke, N.; Joy, Andrew; Hubbins, Stephen; Amerasekera, A.; Yang, C.-K.K., "Adaptation of CDR and full scale range of ADC-based SerDes receiver," *VLSI Circuits, 2009 Symposium on* , vol., no., pp.12,13, 16-18 June 2009

2. B. Schell and Y. Tsividis, 'A Continuous-Time ADC/DSP/DAC System With No Clock and With Activity-Dependent Power Dissipation', *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 2472–2481, 2008.

3. J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, 'Beyond Nyquist: Efficient Sampling of Sparse Bandlimited Signals', *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 520–544, 2010.

4. S. Pfetsch, T. Ragheb, J. Laska, H. Nejati, A. Gilbert, M. Strauss, R. Baraniuk, and Y. Massoud, 'On the feasibility of hardware implementation of sub-Nyquist random-sampling based analog-to-information conversion', *IEEE International Symposium on Circuits and Systems*, 2008.

5. D. J. White, P. E. William, M. W. Hoffman, S. Balkir, and N. Schemm, 'Analog sensing front-end system for harmonic signal classification', *IEEE International Symposium on Circuits and Systems*, 2012.

6. E. J. Candes and M. B. Wakin, 'An Introduction To Compressive Sampling', *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

7. F. Chen, A. P. Chandrakasan, and V. M. Stojanovic, 'Design and Analysis of a Hardware-Efficient Compressed Sensing Architecture for Data Compression in Wireless Sensors', *IEEE Journal of Solid-State Circuits*, vol. 47, no. 3, pp. 744–756, 2012.

8. J. Ramirez, J. M. Gorriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, 'Speech/non-speech discrimination based on contextual information integrated bispectrum LRT', *IEEE Signal Processing Letters*, vol. 13, no. 8, pp. 497–500, 2006.

9. J. Sohn, N. S. Kim, and W. Sung, 'A statistical model-based voice activity detection', *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

10. A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. Tschanz, and V. De, 'A 2.3 nJ/Frame Voice Activity Detector-Based Audio Front-End for Context-Aware System-On-Chip Applications in 32-nm CMOS', *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, 2013.

11. S. Lauwereins, W. Meert, J. Gemmeke, and M. Verhelst, 'Ultra-low-power voice-activity-detector through context- and resource-cost-aware feature selection in decision trees', *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

12. R. Sarpeshkar, 'Analog Versus Digital: Extrapolating from Electronics to Neurobiology', *Neural Computation*, vol. 10, no. 7, pp. 1601–1638, 1998.

13. J. Zhang, Z. Wang, and N. Verma, 'A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion', *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015.

14. S.-Y. Hsu, Y. Ho, P.-Y. Chang, C. Su, and C.-Y. Lee, ‘A 48.6-to-105.2 μW Machine Learning Assisted Cardiac Sensor SoC for Mobile Healthcare Applications’, *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 801–811, 2014.

15. B. Murmann, ‘Digitally assisted data converter design’, *2013 Proceedings of the ESSCIRC (ESSCIRC)*, 2013.

16. K. Badami, S. Lauwereins, W. Meert, and M. Verhelst, ‘Context-aware hierarchical information-sensing in a 6μW 90nm CMOS voice activity detector’, *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015.

17. F. J. Fahy, ‘Measurement of acoustic intensity using the cross-spectral density of two microphone signals’, *The Journal of the Acoustical Society of America*, vol. 62, no. 4, 1977.

18. Knowles:: Microphones.’ [Online]. Available: http://www.knowles.com/eng/Products/Microphones. [Accessed: 30-Apr-2015]

19. Invensense :: Microphones.’ [Online]. Available: http://www.invensense.com/mems/microphone/. [Accessed: 30-Apr-2015]

20. I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, ‘Development and analysis of an International Speech Test Signal (ISTS)’, *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, 2010.

21. S. S Stevens, J. Volkmann, and E. B. Newman, ‘A scale for the measurement of the psychological magnitude pitch’, *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185-190, 1937.

22. S. Lauwereins, K. Badami, W. Meert, and M. Verhelst, ‘Optimal resource usage in ultra-low-power sensor interfaces through context- and resource-cost-aware machine learning’, *Neurocomputing*, 2015.

23. Y. Hu and P. C. Loizou, ‘Subjective comparison and evaluation of speech enhancement algorithms’, *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, 2007.

24. T. Fawcett, ‘An introduction to ROC analysis’, *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

25. Kola, Jonathan, Carol Espy-Wilson, and Tarun Pruthi. "Voice activity detection." MERIT BIEN (2011): 1-6.

26. Yip, M.; Chandrakasan, A.P., "A Resolution-Reconfigurable 5-to-10-Bit 0.4-to-1 V Power Scalable SAR ADC for Sensor Applications," *IEEE Journal of Solid-State Circuits*, vol.48, no.6, pp.1453-1464, 2013

27. B. Rumberg, D. W. Graham, V. Kulathumani, and R. Fernandez, ‘Hibernets: Energy-Efficient Sensor Networks Using Analog Signal Processing’, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 3, pp. 321–334, 2011.

28. S. A. House, ‘On Vowel Duration in English’ *The Journal of the Acoustical Society of America*, vol. 33, pp. 1174-1178, 1961

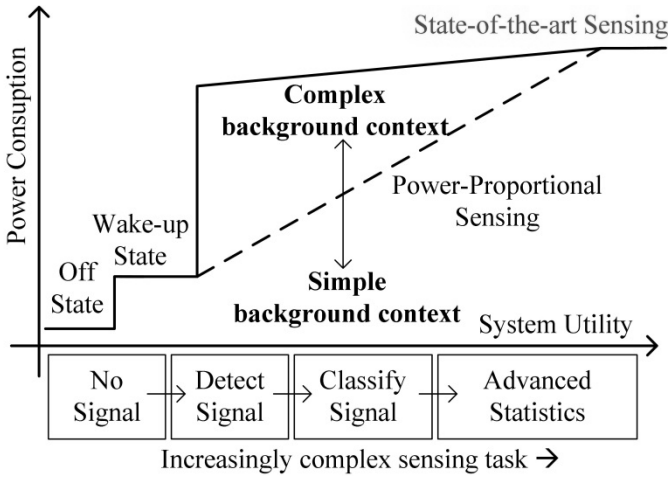

Fig. 1 Power-Proportional sensing in contrast with State-of-the-art sensing systems

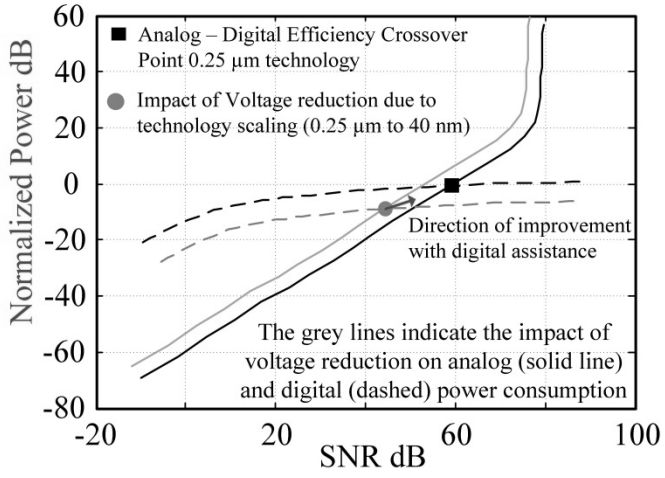

Fig. 2 Computation power scaling for analog (solid line) and digital (dashed line) implementations [12] and impact on efficiency cross over point due to voltage scaling and due to digital assistance by machine learning and / or calibration

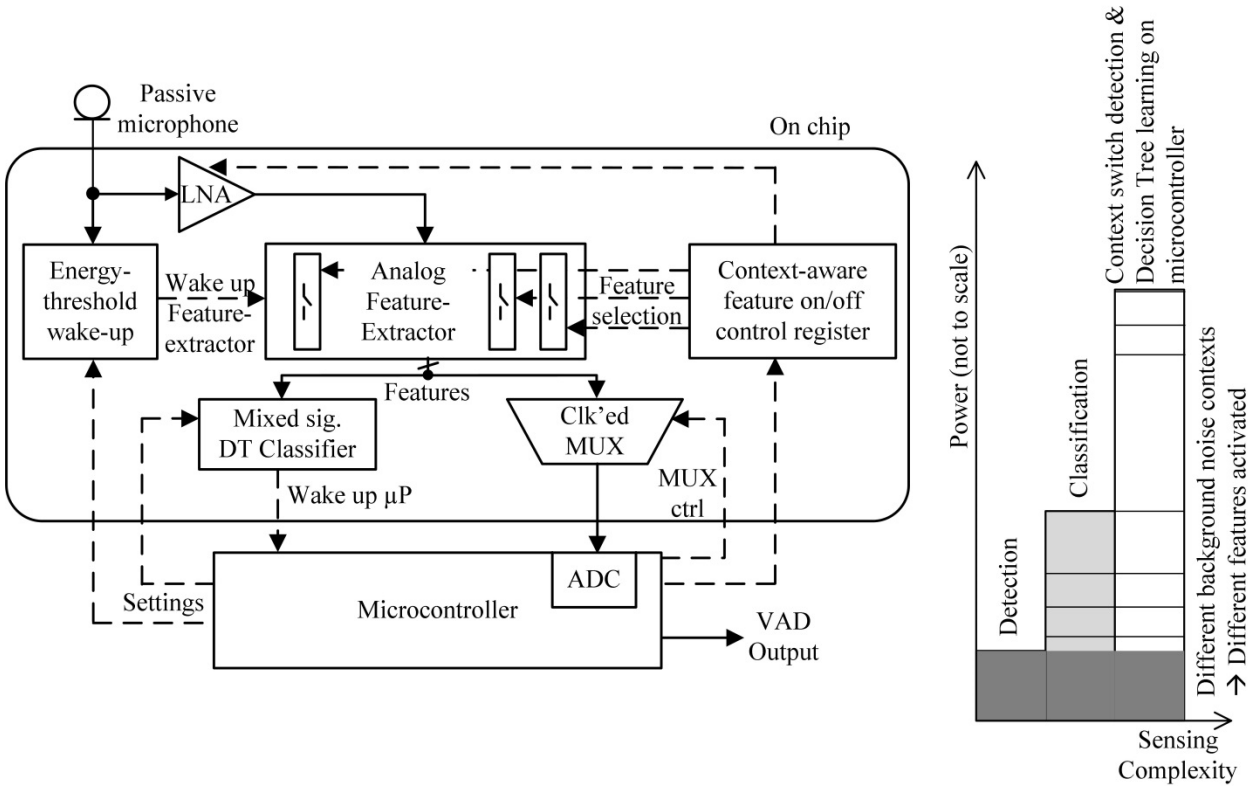

Fig. 3 System diagram of the Power-Proportional voice activity detector (left) and VAD Power scaling with sensing complexity (right)

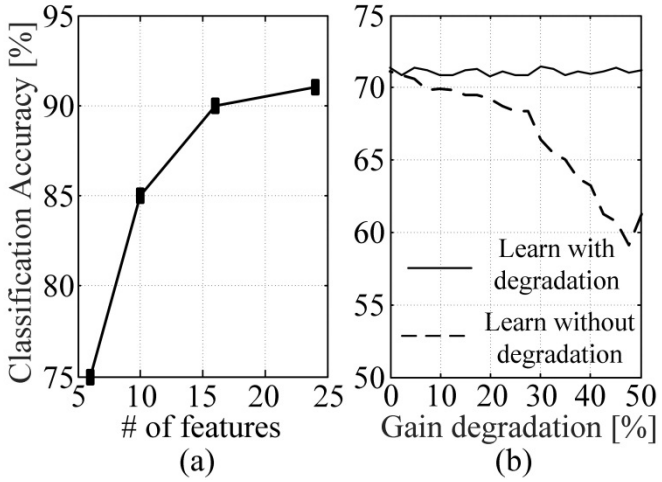

Fig. 4 (a) Impact of number of computed features (b) Impact of gain degradation on classification accuracy. The results are for exhibition background noise with 12dB and 0 dB SANR respectively

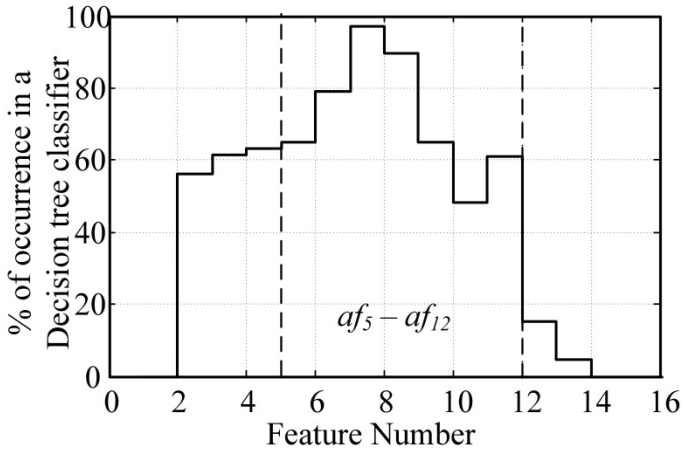

Fig. 5 Histogram depicting average usefulness of computed features in exhibition background noise context for SANR of 0 dB

Table 1 Highlight of important specifications for targeted voice activity detection system.

Frequency Range	Maximum feature count	SANR	Microphone sensitivity	Output Resolution
100 Hz - 5 kHz	16	0 - 12 dB	> -60 dBV	8 bits

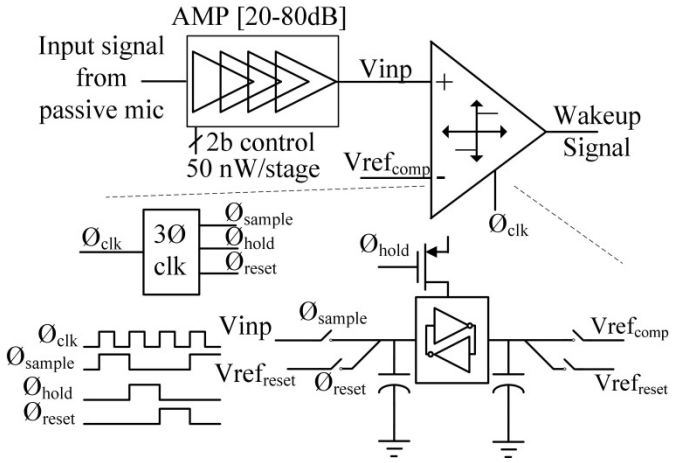

Fig. 6 Schematics for threshold based wakeup detector

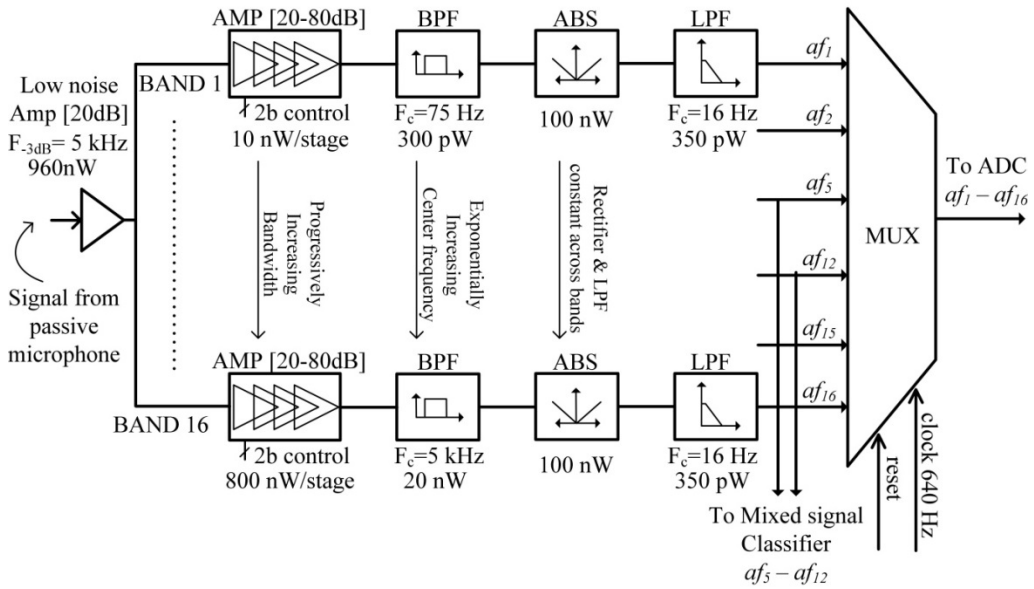

Fig. 7 Schematic and design parameters of the analog feature extraction block

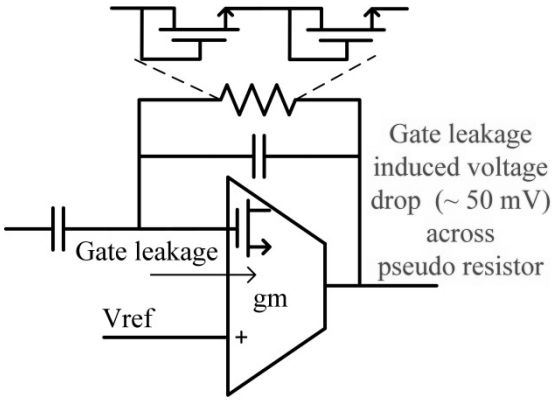

Fig. 8 Amplifier schematic highlighting gate leakage through the input pair

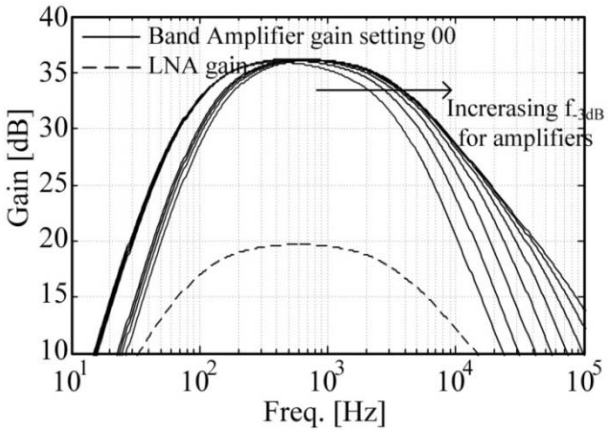

Fig. 9 Simulated frequency response for LNA and amplifiers in even bands showing increasing f_{-3dB}

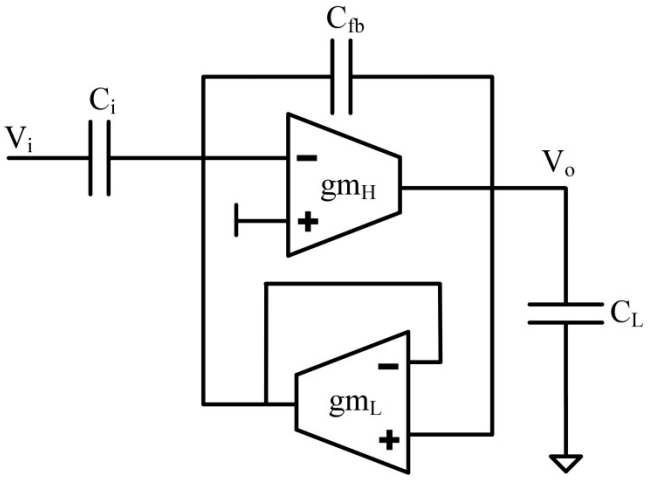

Fig. 10 First order gm – C based band pass filter topology

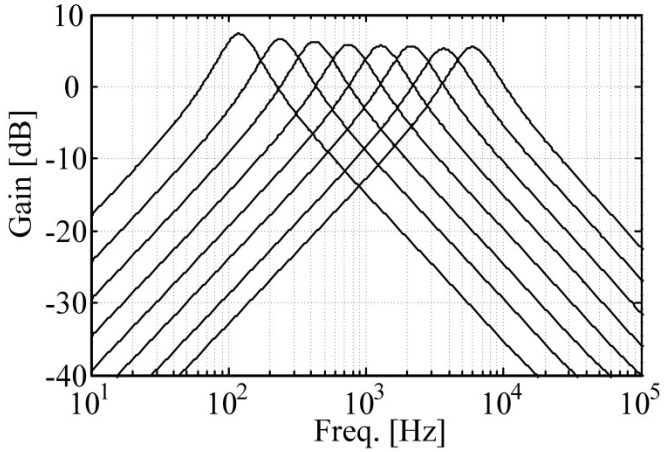

Fig. 11 Simulated frequency response for a constant Q = 1.3 BPF filters in even bands

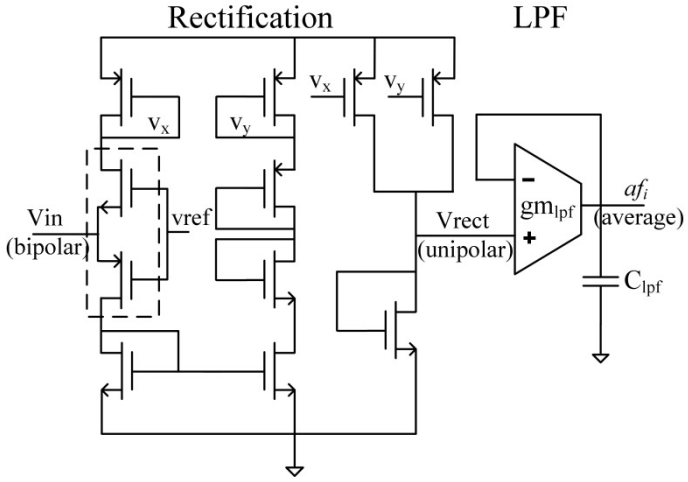

Fig. 12 Rectifier and LPF based averaging circuit

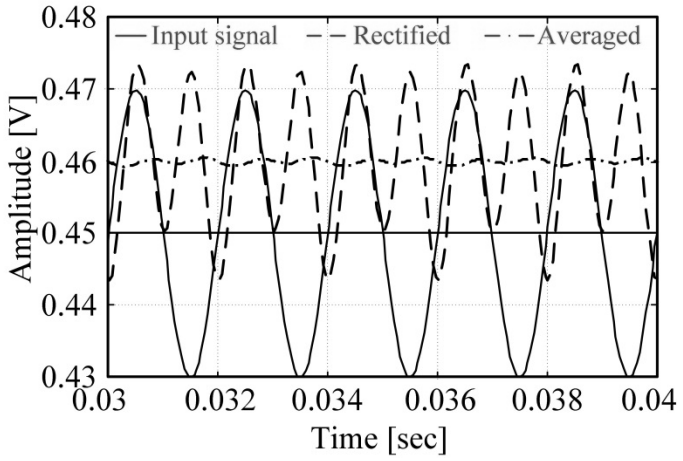

Fig. 13 Simulated response of the averaging circuit for a sinewave input of 20mVpp amplitude and 500 Hz frequency

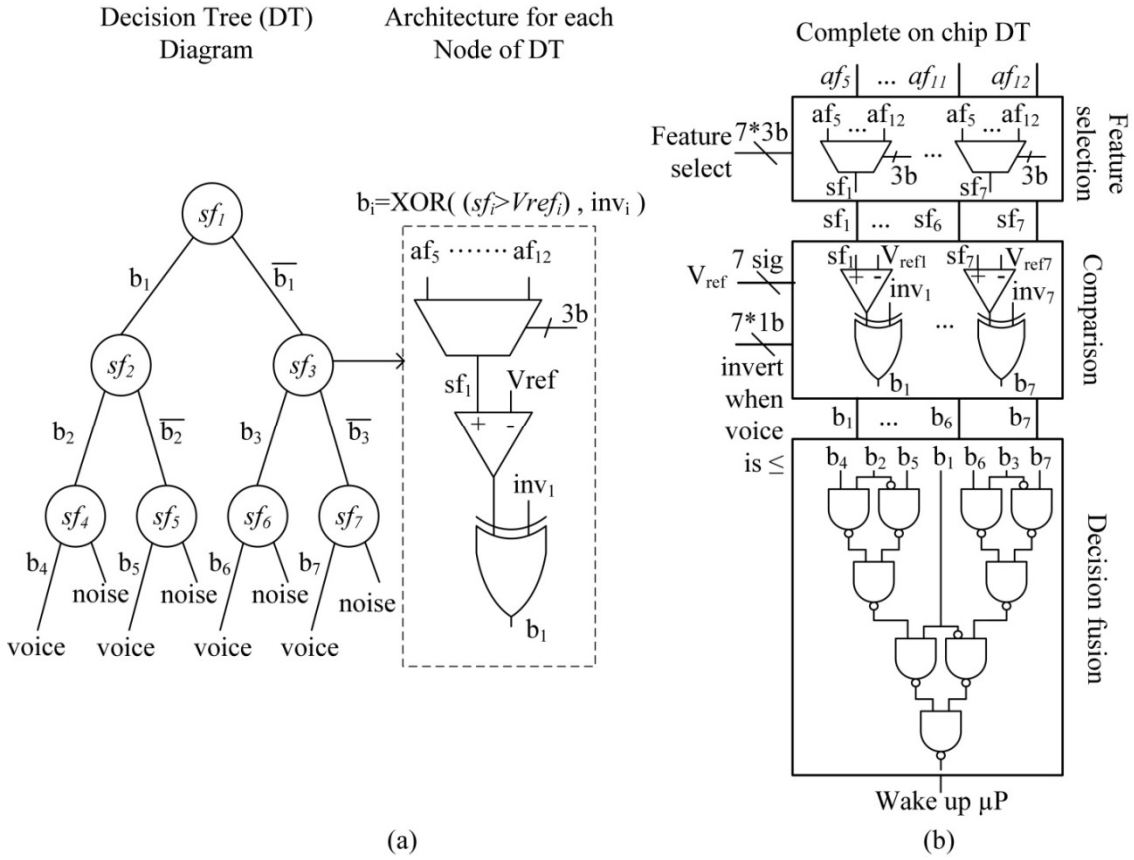

Fig. 14 Architecture of (a) one node of DT classifier and (b) complete classifier

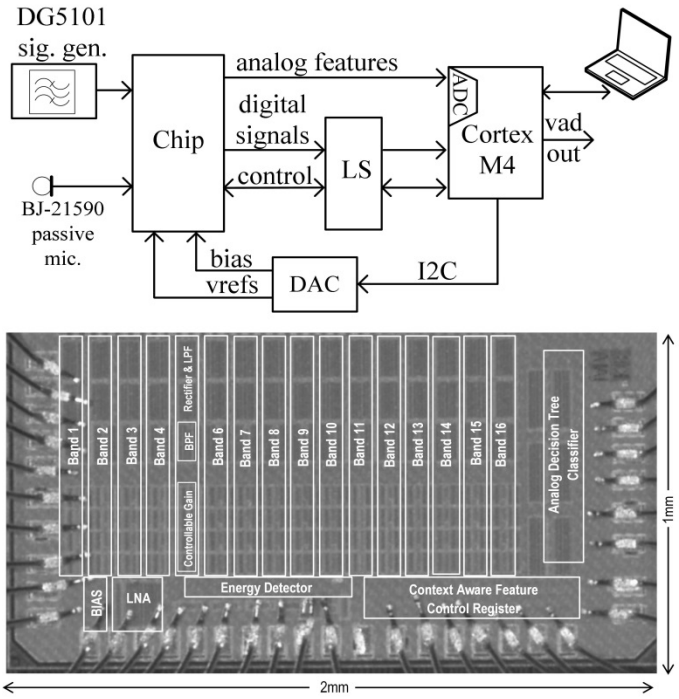

Fig. 15 Measurement setup (top) and chip micrograph (bottom) with important blocks highlighted

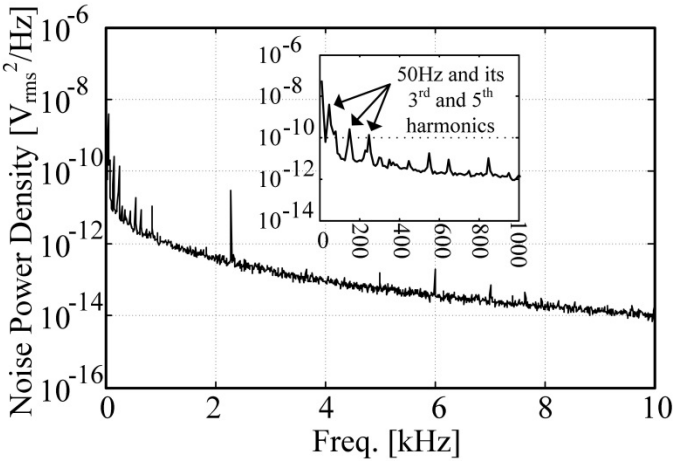

Fig. 16 Measured input referred noise at the LNA output.

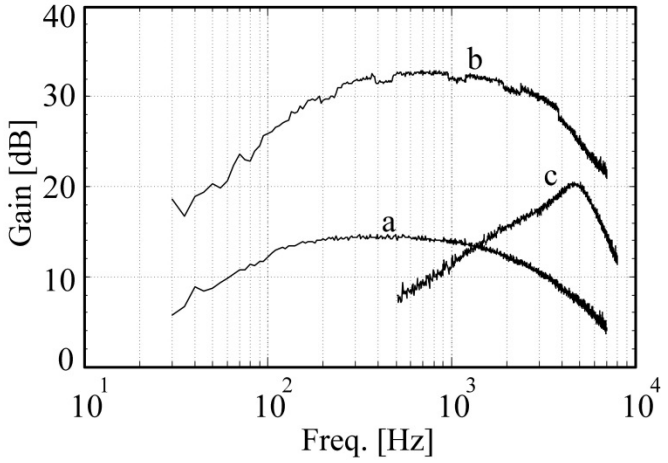

Fig. 17 Measured small signal magnitude response for LNA (a), amplifier with LNA (b), BPF with amplifier (c) in 16^{th} band

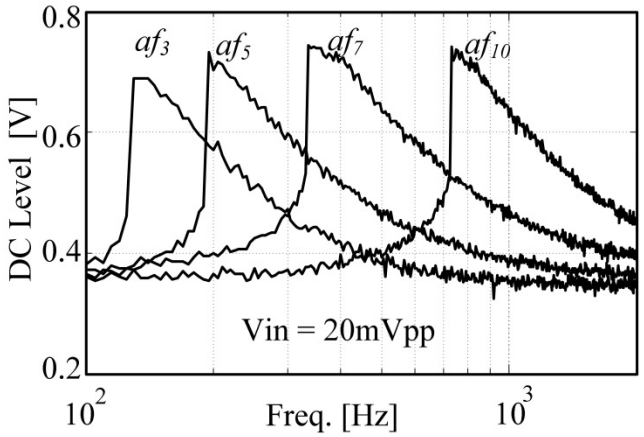

Fig. 18 Measured large signal frequency response of complete bands for bands 3, 5, 7 and 10

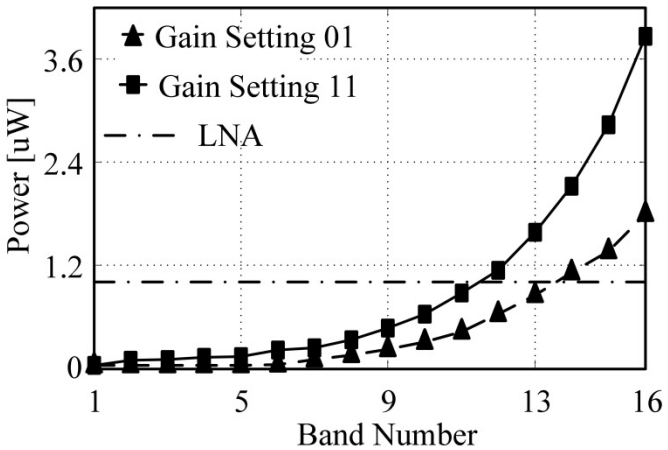

Fig. 19 Measured power consumption of LNA and of each band for gain setting of 01 and 11

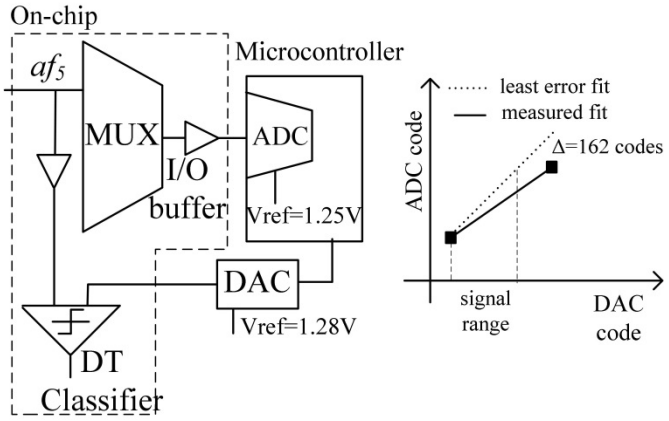

Fig. 20 Calibration scheme for ADC and DAC paths

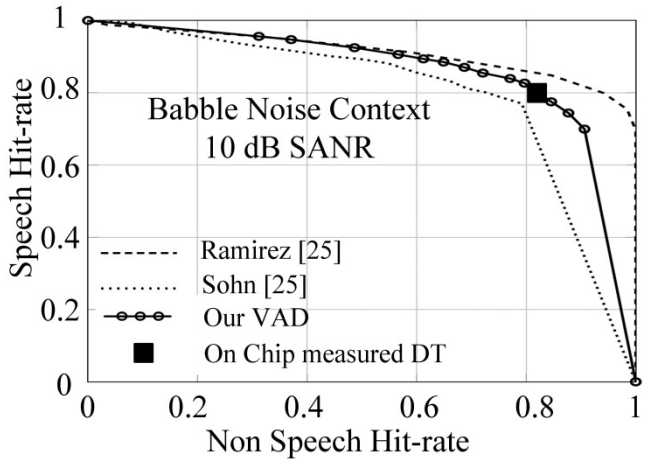

Fig. 21 Comparison of classification accuracy to STOA software VADs

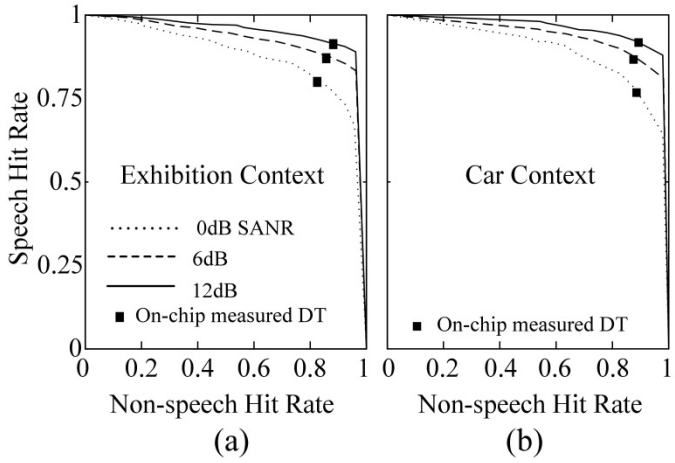

Fig. 22 Measured ROC curves depicting classification accuracy for multiple SANR in (a) Exhibition and (b) Car noise contexts

Table 2 Measured power consumption variation with classification task complexity illustrating achieved Power-Proportional operation

Task	Power
Signal Detect	710 nW
Classify Signal	2.6 µW - Babble Noise 1.1 µW - Exhibition Noise 1.05 µW - Car Noise
Detect Context change and re-learn DT [11]	57 µW
Voice Activity Detection (For babble noise context)	3.8 µW 80% signal detect 15% classification 5% detect context change and retrain

Table 3 Comparison with State of the art VAD and similar systems

	This Work	[27] JETCAS '11	[10] JSSC' 13	[25]
Tech.	90nm CMOS	0.5um CMOS	32nm CMOS	Software only
Area	2mm^2	2.25mm^2	86K gates	NA
Power (feature extraction + classification)	6µW Worst case, all bands on	51µW	< 50µW	>90µW estimated [11]
Gain ncesary for passive mic.	On chip	Off chip	assumes digital mic.	NA
Feature type	Analog	Analog	Digital	Software
Classifier	On chip - Mixed Signal	Off chip - Digital	On chip - Digital	Software based
Context Aware	Yes	NA	Yes	Yes
Feature-Cost aware	Yes	NA	No	No
Latency	< 100ms	100ms	10ms	10ms
Classifier accuracy @ 12dB SNR	HR SP 89% HR Non SP 85% @ Babble 12dB SANR	90% car vs truck classification	97% Unspecified SNR / context / database	HR SP 89% HR Non SP 79% @ Babble 12dB SANR